

Advancing Science through User-Guided Learning in Massive Data Streams

Kirk Borne
George Mason University
kborne@gmu.edu

September 30, 2009

Abstract

Science projects from all disciplines are producing enormous data repositories, which pose both rich targets for exploration and difficult challenges for data mining. An even greater challenge is posed by high-rate data streams from a vast array of sensors and high-efficiency experiments. Mining the knowledge from these data streams is complicated both by the data volume and the time criticality of identifying novel and important events in the data flow. In addition, data have "inertia" – it is easier to analyze the data while they are moving through the cloud than to extract them later from some mega data repository and feed them through memory bottlenecks for offline analysis and mining. For cases where the data flow has no end, the problem gets cumulatively worse with time. Some examples of this will be presented, including astronomy and earth system science: the detection of wildfires in remote sensing networks (for estimating the thermal and carbon inputs to climate change models). We present an intriguing model for user-guided learning from these massive data streams. The approach is analogous to that of particle physics beam experiments, which may produce as much as one petabyte of data per second – these data cannot possibly be stored for later mining and analysis, but they are mined in real-time, since that is the only realistic opportunity to look at the data. The analysis algorithms and cyberinfrastructure that are employed to cope with the particle physics data flood are highly sophisticated. We are investigating another approach – human computation – which is characterized by enormous cognitive capacity and pattern recognition efficiency. We will describe some remarkable results from the field of citizen science, based upon work with static databases. We envision the eventual application of this emerging computational resource to the problem of massive stream mining for scientific discovery.